

# Power-law relations in random networks with communities

Clara Stegehuis, Remco van der Hofstad, and Johan S.H. van Leeuwen  
*Eindhoven University of Technology, Department of Mathematics and Computer Science,  
P.O. Box 513, 5600 MB Eindhoven, The Netherlands*

Most random graph models are locally tree-like – do not contain short cycles – rendering them unfit for modeling networks with a community structure. We introduce the hierarchical configuration model (HCM), a generalization of the configuration model that includes community structures, while properties such as the size of the giant component, and the size of the giant percolating cluster under bond percolation can still be derived analytically. Viewing real-world networks as realizations of HCM, we observe two previously undiscovered power-law relations: between the number of edges inside a community and the community sizes, and between the number of edges going out of a community and the community sizes. We also relate the power-law exponent  $\tau$  of the degree distribution with the power-law exponent of the community size distribution  $\gamma$ . In the case of extremely dense communities (e.g., complete graphs), this relation takes the simple form  $\tau = \gamma - 1$ .

PACS numbers: 64.60.aq, 89.75.-k, 64.60.ah

## I. INTRODUCTION

Random graphs serve to model large networked systems, but are typically unfit for modeling community structure. Communities refer to relatively densely connected groups of vertices, with more sparse connections between groups, and the community structure refers to an arbitrary number of groups, each of arbitrary size and structure. In this paper we introduce the Hierarchical Configuration Model (HCM), a model for generating networks as random graphs with not only arbitrary degree distribution but also arbitrary community structure. The HCM is directly applicable to network data, remains solvable in the large-network limit for properties related to component sizes, clustering coefficients and percolation, and is a natural extension of the widely studied configuration model (CM) [8, 28, 33] for random graphs with a given degree distribution.

Communities are relatively densely connected and contain relatively many short cycles. Since the CM contains only few short cycles, it cannot model networks with community structure. One possibility to add community structure to random graphs is by adding so-called households [2, 3, 14, 38]. In this line of work, on the macroscopic level, the graph is initially a CM in which each vertex of the graph can be replaced by a complete graph (referred to as household). Vertices in a household have links to all other household members, which creates a community structure. These household models allow to study networks with a prescribed degree distribution and a tunable clustering coefficient, because the clustering coefficient can be manipulated by the household structure. Hence, the focus in [2, 3, 14, 38] is on locally incorporating short cycles to explain clustering at the global network level. In a similar spirit, a class of random graphs was introduced in [30] in the form of a random network that only contains random edges and triangles. Each vertex is assigned the number of triangles it is in. The triangles are formed by pairing the nodes

at random, and regular edges are formed according to the statistical rules of the CM. The model in [30] was extended in [23] to networks with arbitrary distributions of subgraphs.

Like in these previous works, our goal is to develop a more realistic yet tractable random network model, by creating conditions under which the tree-like structure is violated within the communities, but remains to hold at a higher network level – the network of communities in our case. There are, however, considerable differences with these earlier works. The model in [23, 30] departs from a specification of all possible subgraphs or motifs, which is the triangle in [30] and all possible subgraphs in [23]. The network is then created by specifying the number of subgraphs attached to each vertex and then sampling randomly from the set of compatible networks. A community can thus exist of many subgraphs, think of a large cluster of triangles, which makes the framework in [23, 30] harder to fit on real-world networks. In fact, in [23] the appropriate selection of subgraphs and their frequencies for practical purposes is mentioned as a challenging open problem. The approaches in [2, 3, 14, 30, 38] are geared towards increasing clustering and fitting a global clustering coefficient, but are less suitable to directly describe community structure. Like [14, 38] we construct a random graph model that at the higher level is a tree-like configuration model, and at the lower level contains subgraphs, but these subgraphs do not need to be complete graphs. Large real-world communities are relatively dense, but not necessarily *completely* connected. We thus generalize the setting of [2, 14, 38] to arbitrary community structures, to account for heterogeneity in size and internal connectivity.

The HCM breaks away from previous models with clustering or communities, because the model can use any proposed community structure as input. That is, the HCM viewed as an algorithm first models the community structure, and only then creates the random network model. This top-down approach is in sharp contrast with

the bottom-up approach taken in [2, 3, 14, 30, 38]. To be more specific, the community structure can be uncovered by some detection algorithm that, when applied to a real-world network, leads to a collection of plausible communities and their frequencies. By sampling from this collection of communities, the HCM can generate resampled networks with similar structure as the original network. The HCM thus enriches standard random graph models with the ability to describe random yet realistic community structures. Where the CM is the canonical model for complex networks with power-law degree sequences, the HCM adds to this the community structure.

The main contribution of this paper is the introduction and analysis of the HCM. As is common for the CM [33], we perform our study under the assumption of locally tree-like approximations, ignoring the presence of double edges and cycles. Using this *tree ansatz* we can apply the generating function formalism [33] to obtain analytical results. A fully rigorous mathematical treatment of the HCM, taking multiple edges and self-loops into account, is performed in a companion paper [21]. Based on our analysis of the HCM, we discuss several phenomena that each will spark off future research directions.

Our work reveals a potentially crucial property of real-world networks that has received virtually no attention: the joint distribution  $p_{k,s}$  of the community size  $s$  and the number of connections  $k$  a community has to other communities. The size of the giant component delicately depends on this joint distribution, which can be determined from network data once the community structure is determined. In fact, after studying  $p_{k,s}$  for several real-world networks, we observe a *power-law relation between the size of the communities and the number of edges out of a community in many real-world networks*.

Except for this joint distribution, the size of the giant component does not depend on detailed information about the structure of the communities. When we perform percolation on the network, more precise structural information does matter. To see this, imagine a process spreading over the network, by starting at some vertex and traversing according to some rule to all vertices in the connected component to which this vertex belongs. Before percolation, once the process reaches a vertex in a community, it reaches the entire community. But after percolation, the community no longer needs to be connected, so that parts of the community may become unreachable. Despite this difficulty, we are able to describe the percolation phase transition explicitly. Inspired by this need to include detailed community structure, and thus extend the model description beyond  $p_{k,s}$ , we observe a second *power-law relation between the denseness of a community and its size* in several real-world networks.

For the present paper, the most important application of the HCM is to investigate power-law networks. Statistical analysis of network data shows that many networks possess a power-law degree distribution [12, 31, 32, 39]. Traditionally, this is captured by using the CM and as-

suming that the probability  $p_k$  that a node has  $k$  neighbors then scales with  $k$  as  $p_k \sim Ck^{-\tau}$  for some constant  $C$  and power-law exponent  $\tau > 0$ . Many scale-free networks have an exponent  $\tau$  between 2 and 3 [1, 4, 15], so that the second moment of the degree distribution diverges in the infinite-size network limit. The exponent  $\tau$  is an important characteristic of the network and determines for example the mean distances in the network [19, 20, 33], or the behavior of various processes on the graph like bond percolation [9], first passage percolation [5] and an intentional attack [13]. Using the HCM instead of the CM, it no longer suffices to describe the degree distribution  $p_k$ , but instead assumptions need to be made about the joint distribution  $p_{k,s}$ . In the special case of extremely dense communities, this joint perspective gives rise to the following phenomenon (that we formalize in Section IV): *If the total degree distribution of a network with extremely dense communities follows a power law with exponent  $\tau$ , the power law of the community sizes has exponent  $\gamma = \tau + 1$ .* In the household model, where each community is a complete graph, we observe that indeed  $\gamma = \tau + 1$ . However, real-world network data shows that communities are not always extremely dense, in which case we find that  $\gamma \neq \tau + 1$ . This is due to the fact that the edge density of communities turns out to decay with community size.

The outline of this paper is as follows. In Section II we introduce the model and show how generating function techniques can be used to obtain exact large-graph limit results for the giant component. In Section III we consider percolation on the HCM and again using generating function techniques we obtain the critical point and the size of a giant percolation cluster. We further investigate the delicate relation between community structure and percolation and show that community structure can both enforce and inhibit percolation. In Section IV we consider the special case in which the degree distribution follows a power law. It is here that we discover the power-law shift caused by community structure when the communities are *extremely dense*. In Section V we apply the HCM to several real-world networks, and we observe two more power-law relations in graphs with communities. In Section VI we present conclusions and future research directions.

## II. MODEL DESCRIPTION

We define the HCM as an extension of the CM, in which each vertex in the CM is replaced by a community – some connected graph. We denote community  $i$  by  $H_i = (F_i, \mathbf{d}_i)$ . Here  $F_i = (V_{H_i}, E_{H_i})$  is a graph, defining the shape of the community. The vector  $\mathbf{d}_i$  counts the number of half-edges attached to each vertex in  $F_i$ , going out of the community. These half-edges will form the inter-community connections. Let  $s_i$  be the size of community  $i$ , and  $k_i$  the number of edges from community  $i$  to other communities. Figure 1 shows an example

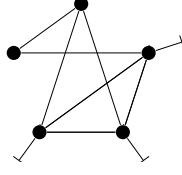


Figure 1. A community with  $s = 5$  and  $k = 3$

of a community with  $s = 5$  and  $k = 3$ . If we order the vertices clockwise, starting at the vertex in the top of the graph, then  $\mathbf{d} = (0, 1, 1, 1, 0)$ . The number of half-edges attached to vertex  $v$  is denoted by  $d_v^{(b)}$ , and referred to as *outside degrees* or inter-community degrees, which define the connections between communities. Similarly, the number of edges from vertex  $v$  to other vertices inside the same community is denoted by  $d_v^{(c)}$ , the *inside degrees* or intra-community degrees, the collection of which defines the local connections inside the communities. The degree of vertex  $v$  satisfies  $d_v = d_v^{(b)} + d_v^{(c)}$ . As in the CM, the random graph is constructed by picking two half-edges at random, and pairing them. This procedure continues until no half-edges are left. Thus, the edges that connect different communities are formed as in the CM, but the intra-community edges are fixed.

We define the joint distribution  $p_{k,s}$  to be the fraction of communities of size  $s$  having outside degree  $k$ . We denote the number of vertices in the random graph by  $N$  and the number of communities by  $n$ . To calculate properties of the random graph, we define several distributions and their probability generating functions (pgfs). Let

$$g_p(x, y) = \sum_{k,s} p_{k,s} x^k y^s \quad (1)$$

denote the pgf of the joint size and outside degree distribution of the communities. Introduce the *excess outside degree distribution* by

$$q_{k,s} = \frac{(k+1)p_{k+1,s}}{\langle k \rangle}, \quad (2)$$

where  $q_{k,s}$  can be interpreted as the probability to arrive in a community of size  $s$  and outside degree  $k+1$  when traversing a random inter-community edge, including the traversed edge. Here  $\langle k \rangle = \sum_{k,s} k p_{k,s}$  is the expected value of  $k$ . Similarly, define

$$r_{k,s} = \frac{s p_{k,s}}{\langle s \rangle} \quad (3)$$

as the probability that a randomly chosen vertex is in a community of size  $s$  (including the vertex itself) and has  $k$  edges to other communities. The pgfs for these

probability distributions are given by

$$\begin{aligned} g_q(x, y) &= \sum_{k,s} q_{k,s} x^k y^s = \frac{1}{\langle k \rangle} \sum_{k,s} k p_{k,s} x^{k-1} y^s \\ &= \frac{1}{\langle k \rangle} \frac{\partial g_p(x, y)}{\partial x}, \end{aligned} \quad (4)$$

$$\begin{aligned} g_r(x, y) &= \sum_{k,s} r_{k,s} x^k y^s = \frac{1}{\langle s \rangle} \sum_{k,s} s p_{k,s} x^k y^{s-1} \\ &= \frac{y}{\langle s \rangle} \frac{\partial g_p(x, y)}{\partial y}. \end{aligned} \quad (5)$$

We use these pgfs to calculate the size of the largest connected component in the graph.

### A. Emergence and size of a giant component

Let us first explain why the HCM remains amenable for analysis using the generating function technique. Although the HCM is not locally tree-like, the connections between communities are formed as in the CM. Therefore, on the higher level of communities, the HCM is still locally tree-like, and the probability that a half-edge attached to a community forms a self-loop or multiple edge with other communities, tends to zero as  $N$  grows large.

Let  $u$  be the probability that a community that is reached by traversing a random inter-community edge is not in the giant component, in which case all the communities connected to it cannot be in the giant component either. The  $k$  neighboring communities of the reached community are not in the giant component with probability  $u^k$ . Hence, a community is not in the giant component with probability

$$u = \sum_{k,s} q_{k,s} u^k = g_q(u, 1). \quad (6)$$

Similarly, the probability that a randomly chosen vertex is not in the giant component is  $\sum_{k,s} r_{k,s} u^k = g_r(u, 1)$ . Thus, the size of the largest component  $S$  satisfies

$$S = 1 - g_r(u, 1). \quad (7)$$

A giant component exists if and only if  $S > 0$ . We can see that  $S = 0$  if and only if  $u = 1$ , so  $S > 0$  for  $u < 1$ . Furthermore, (6) shows that  $u$  is the extinction probability of a branching process with degree distribution  $\sum_s q_{k,s}$  and expected offspring

$$\sum_{k,s} k q_{k,s} = \frac{\langle k(k-1) \rangle}{\langle k \rangle}. \quad (8)$$

It is well known that the condition  $u < 1$  for the existence of a giant component is equivalent to the expected offspring being larger than one, or  $\langle k^2 \rangle - 2\langle k \rangle > 0$  [28]. This is the same condition as for the CM with offspring  $(p_k)_{k \geq 0}$ , so that the point at which the giant component

emerges is only determined by the inter-community connections. Here we have made the implicit assumption that  $\langle s \rangle < \infty$ , to be able to use (5). As long as  $\langle s \rangle < \infty$ , the community sizes have no influence on the point at which the giant component emerges.

In general,  $k$  and  $s$  can be dependent. If  $k$  and  $s$  were independent, that is  $p_{k,s} = p_k p_s$  and

$$u = \langle k \rangle^{-1} \sum_{k,s} k p_k p_s u^{k-1} = \langle k \rangle^{-1} \sum_k k p_k u^{k-1}, \quad (9)$$

then

$$S = 1 - \langle s \rangle^{-1} \sum_{k,s} s p_k p_s u^k = 1 - \sum_k p_k u^k. \quad (10)$$

These equations are the same as for the CM with degree distribution  $(p_k)_{k \geq 0}$ . Therefore, when  $k$  and  $s$  are independent, also the size of the giant component is entirely defined by the inter-community connections. However, in real-world networks,  $k$  and  $s$  are likely to be dependent. Independence would imply that a small community has the same probability of having a large number of edges towards other communities as a large community. In most real-world examples it is more likely that every vertex inside a community has some edges towards other communities, so a larger community has a larger probability of  $k$  being large than a smaller community, in which case  $k$  and  $s$  are positively correlated.

It is also possible to calculate the sizes of the other components, when there is no giant component. Let  $h_q(z)$  be the pgf of the number of vertices accessible from a randomly chosen inter-community edge. When the edge reaches a community of size  $s$ , this adds  $s$  vertices to the component, which contributes a factor  $z^s$ . Furthermore, if the community has  $k$  other outgoing edges, then each of these edges will generate a component with pgf  $h_q(z)$ . This gives

$$h_q(z) = \sum_{k,s} q_{k,s} z^s h_q(z)^k = g_q(h_q(z), z). \quad (11)$$

Now we derive  $h_p(z)$ , the pgf of the size of the component of a uniformly chosen vertex. When a uniformly chosen vertex is in a community of size  $s$  and outside degree  $k$ , the members of the community add  $z^s$  to the pgf. Each half-edge generates a component with pgf  $h_q(z)$ . This gives

$$h_p(z) = \sum_{k,s} r_{k,s} z^s h_q(z)^k = g_r(h_q(z), z). \quad (12)$$

The mean component size is given by  $h'_p(1)$ . Differenti-

ating (11) and (12) yields

$$\begin{aligned} h'_q(1) &= \frac{1}{\langle k \rangle} \frac{\partial g_p}{\partial x^2}(1,1) h'_q(1) + \frac{1}{\langle k \rangle} \frac{\partial g_q}{\partial xy}(1,1) \\ &= \frac{\langle k(k-1) \rangle}{\langle k \rangle} h'_q(1) + \frac{\langle ks \rangle}{\langle s \rangle}, \end{aligned} \quad (13)$$

$$\begin{aligned} h'_p(1) &= \frac{\partial g_r}{\partial x}(1,1) h'_q(1) + \frac{\partial g_r}{\partial y}(1,1) \\ &= \frac{\langle ks \rangle}{\langle s \rangle} h'_q(1) + \frac{\langle s^2 \rangle}{\langle s \rangle}. \end{aligned} \quad (14)$$

These equations together define  $h'_q(1)$  and  $h'_p(1)$  and some rewriting yields

$$h'_q(1) = \frac{\langle ks \rangle}{2 \langle k \rangle - \langle k^2 \rangle}, \quad (15)$$

$$h'_p(1) = \frac{\langle s^2 \rangle}{\langle s \rangle} + \frac{\langle ks \rangle^2}{\langle s \rangle (2 \langle k \rangle - \langle k^2 \rangle)}. \quad (16)$$

when  $2 \langle k \rangle - \langle k^2 \rangle > 0$ . The first term in (16) is the expected size of the component in which the randomly chosen vertex lies. The second term equals the expected size of the components attached to the community of the randomly chosen vertex. The expected component size is infinite if the giant component emerges (if  $2 \langle k \rangle - \langle k^2 \rangle > 0$ ). Equation (16) shows that  $\langle ks \rangle = \infty$  or  $\langle s^2 \rangle = \infty$  also give an infinite expected component size. However, the condition for the giant component to emerge does not involve  $s$ . Thus, it is possible to have an infinite expected cluster size without having a giant component. Then the expected cluster size is infinite, but still small compared to the total number of vertices in the graph. One example of a random graph with no giant component but an infinite expected cluster size is a graph with  $k = 0$  for all communities (all communities are isolated), and  $p_s = C s^{-\alpha}$  for some  $\alpha \in (2, 3)$ . This example has  $\langle s^2 \rangle = \infty$ , hence by (16) the expected cluster size of a randomly chosen vertex is infinite. Since  $k = 0$  for all communities, clearly condition (8) is not met, and there is no giant component. The size of the largest component in this example is the size of the largest community  $s_{\max}$ . If there are  $n$  communities, then  $s_{\max} \sim n^{1/(\alpha-1)}$  [29]. The total number of vertices in the HCM with  $n$  communities goes to  $n \langle s \rangle$ . Thus, the fraction of vertices in the largest component behaves like

$$\frac{n^{1/(\alpha-1)}}{n \langle s \rangle} = \frac{1}{\langle s \rangle} n^{\frac{2-\alpha}{\alpha-1}} \rightarrow 0, \quad (17)$$

as  $n \rightarrow \infty$ . The same phenomenon occurs when the community size distribution is the same, but we add some edges between communities. So as long as  $\langle k^2 \rangle - 2 \langle k \rangle < 0$ , there is no giant component, even though the expected cluster size is infinite.

### III. PERCOLATION

In this section we consider bond percolation on the HCM, where each edge is removed independently with probability  $1 - \phi$ . We are interested in the size of the giant component after removing the edges. We calculate this in a similar way as we computed the size of the giant component before percolation. Next to the joint distribution  $p_{k,s}$  we define the distribution  $p_H$  that denotes the fraction of communities of type  $H$ .

Deleting each edge with probability  $1 - \phi$  is the same as first deleting only the intra-community edges with probability  $1 - \phi$  and then the inter-community edges also with probability  $1 - \phi$ . Thus, first delete each edge inside the communities with probability  $1 - \phi$ . After this procedure, some communities may have split into several connected components. However, these components still form connections as in the CM. Hence, after percolation inside the communities, we again have an HCM. The communities in the new HCM are the connected components of the percolated communities.

When entering a percolated community, the percolated community no longer needs to be connected, so that the  $k$  edges to other communities are not always reached. To account for this, we introduce  $f(H, v, l, \phi)$ , the probability that after percolation, the connected component of community  $H$  containing vertex  $v$  still has  $l$  outgoing edges. Let  $t_k^{(\phi)}$  be the probability that a randomly chosen vertex is in a percolated community with  $k$  edges to other communities. Vertex  $v$  in community  $H$  is chosen with probability  $p_H/s_H$ . The probability that  $v$  is connected to  $k$  half-edges is given by  $f(H, v, k, \phi)$ . Hence,  $t_k^{(\phi)}$  is given by

$$t_k^{(\phi)} = \sum_H \sum_{v \in V_H} \frac{1}{s_H} p_H f(H, v, k, \phi). \quad (18)$$

Let  $q_k^{(\phi)}$  denote the probability that a percolated community reached by traversing a randomly chosen inter-community edge has  $k$  edges towards other communities. The probability of arriving in vertex  $v$  of community  $H$  is proportional to  $p_H d_v^{(b)}$ . Then the probability that  $v$  is inside a percolated community with  $k$  other outgoing edges is  $f(H, v, k + 1, \phi)$ , so that

$$q_k^{(\phi)} \propto \sum_H \sum_{v \in V_H} d_v^{(b)} p_H f(H, v, k + 1, \phi). \quad (19)$$

This gives

$$\begin{aligned} q_k^{(\phi)} &= \frac{\sum_H \sum_{v \in V_H} d_v^{(b)} p_H f(H, v, k + 1, \phi)}{\sum_l \sum_H \sum_{v \in V_H} d_v^{(b)} p_H f(H, v, l, \phi)} \\ &= \frac{\sum_H \sum_{v \in V_H} d_v^{(b)} p_H f(H, v, k + 1, \phi)}{\langle k \rangle}. \end{aligned} \quad (20)$$

Define  $g_{q^{(\phi)}}(z)$  and  $g_{t^{(\phi)}}(z)$  as the pgf of  $q_k^{(\phi)}$  and  $t_k^{(\phi)}$  respectively, which will be used to calculate the size of the giant percolation cluster.

#### A. Giant percolation cluster

After percolating the edges inside communities, we percolate the edges between communities. Since the inter-community edges are paired as in the CM, percolation on these edges is similar to percolation on the CM [9, 22]. We remove each half-edge of a community with probability  $1 - \sqrt{\phi}$ . Then an edge is removed when at least one of the two half-edges that form the edge is removed. Thus, an edge is removed with probability  $2(1 - \sqrt{\phi})\sqrt{\phi} + (1 - \sqrt{\phi})^2 = 1 - \phi$ , as required.

Given that the number of half-edges that are attached to a community before percolating is  $k$ , the number of half-edges after percolating has a binomial distribution with parameters  $(k, \sqrt{\phi})$ . If we denote by  $X^{(\phi)}$  and  $Q^{(\phi)}$  the number of half-edges of a community entered via a randomly chosen edge after and before percolation, respectively, then

$$\begin{aligned} g_{X^{(\phi)}}(z) &= \sum_{l=1}^{\infty} z^l \mathbb{P}(X^{(\phi)} = l) \\ &= \sum_{l=1}^{\infty} \sum_{k=1}^{\infty} z^l \mathbb{P}(X^{(\phi)} = l \mid Q^{(\phi)} = k) q_k^{(\phi)} \\ &= \sum_{k=1}^{\infty} q_k^{(\phi)} \sum_{l=1}^{\infty} z^l \mathbb{P}(\text{Bin}(k, \sqrt{\phi}) = l) \\ &= \sum_{k=1}^{\infty} q_k^{(\phi)} (1 - \sqrt{\phi} + \sqrt{\phi}z)^k \\ &= g_{q^{(\phi)}}(1 - \sqrt{\phi} + \sqrt{\phi}z). \end{aligned} \quad (21)$$

Here  $\mathbb{P}(\text{Bin}(n, p) = i) = \binom{n}{i} p^i (1 - p)^{n-i}$  is the probability that a binomial random variable with parameters  $(n, p)$  takes the value  $i$ .

As in the previous section, let  $u^{(\phi)}$  denote the probability that a vertex reached from a uniformly chosen half-edge is not in the giant component after percolation. One possibility is that the half-edge that we follow links to a deleted half-edge, which happens with probability  $1 - \sqrt{\phi}$ . In this case, the half-edge does not lead to the giant component with probability one. If the chosen half-edge links to a half-edge that was not deleted (which happens with probability  $\sqrt{\phi}$ ), then it leads to a vertex inside a percolated community. This community is not in the giant component if all of its half-edges do not link to the giant component, which happens with probability  $g_{X^{(\phi)}}(u^{(\phi)})$ . This results in

$$u^{(\phi)} = (1 - \sqrt{\phi}) + \sqrt{\phi} g_{q^{(\phi)}}(1 - \sqrt{\phi} + \sqrt{\phi} u^{(\phi)}). \quad (22)$$

A randomly chosen vertex is not in the giant component when all the half-edges of its percolated community do not link to the giant component. Hence, the size of the giant component after percolation  $S^{(\phi)}$  satisfies

$$1 - S^{(\phi)} = g_{t^{(\phi)}}(1 - \sqrt{\phi} + \sqrt{\phi} u^{(\phi)}). \quad (23)$$

Solving equations (22) and (23) together gives the size of the giant component after percolation.

## B. Percolation transition point

To find the percolation transition point, we view the number of communities that can be reached by traversing a random inter-community edge (excluding the traversed edge) as a branching process. The offspring distribution of this branching process is the distribution of the number of half-edges attached to a percolated community reached by traversing a random edge. The expected number of such half-edges after percolation inside the communities is  $\langle q^{(\phi)} \rangle$ . When we then delete the inter-community edges with probability  $1 - \phi$ , the mean number of half-edges of a community reached by traversing a random edge (excluding the traversed edge) is  $\phi \langle q^{(\phi)} \rangle$ . Hence, we view the number of communities that can be reached from a random edge as a branching process with mean offspring  $\phi \langle q^{(\phi)} \rangle$ . This immediately shows that the percolation transition point  $\phi_c$  is when  $\phi_c \langle q^{(\phi_c)} \rangle = 1$ , so that

$$\phi_c = \frac{\langle k \rangle}{\sum_H \sum_{v \in V_H} \sum_k k d_v^{(b)} p_H f(H, v, k+1, \phi_c)}. \quad (24)$$

Equations (23) and (24) show that the critical percolation value as well as the size of the giant component after percolation depend on the shapes of the communities. Appendix A shows that introducing a community structure may either increase or decrease the critical percolation value as well as the size of the giant component after percolation, and can even lead to non-convex percolation curves.

An interesting example is the case of highly connected communities. These communities are robust against percolation compared to the inter-community connections. Then, close to the critical value  $\phi_c$ , these communities will still be connected after percolation, and  $f(H, v, l, \phi) \approx 1$  for  $l = k$ . Then (24) reduces to  $\phi_c = \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$ , as in the standard CM. Thus, if communities are highly connected compared to the inter-community edges, the critical percolation value is entirely determined by the inter-community edges.

## IV. POWER-LAW COMMUNITY SIZES

A potentially crucial property observed in many networks is that the community size distribution appears to have a power-law form over some significant range [7, 11, 18, 35]. We now assume that both the degree and the community size distributions obey power laws with exponents  $\tau$  and  $\gamma$ , respectively. Typical values reported in the literature are  $2 \leq \tau \leq 3$  and  $1 \leq \gamma \leq 3$  [16].

*a. Household communities.* An extreme community structure is that of household communities [38], in which all communities are complete graphs. Each vertex inside the community has outside degree one. Figure 2 shows an example of a household community with  $s = 5$ .

In a household community,  $k = s$ , hence  $p_{k,s} = 0$  if  $k \neq s$ . Suppose the distribution of the community sizes

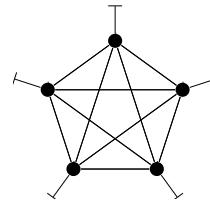


Figure 2. A household community with  $s = 5$

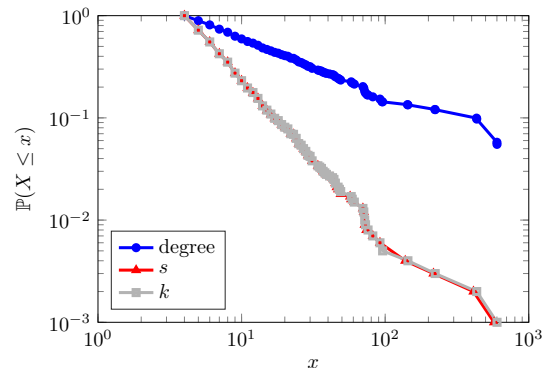


Figure 3. The degree distribution of a household model follows a power law with a smaller exponent than the community size distribution and outside degree distribution

follows a power law with exponent  $\gamma$ ,  $p_{k,k} = Ck^{-\gamma}$ . Then the outside degrees also follow a power-law distribution with exponent  $\gamma$ . Now we derive  $(\hat{p}_k)_{k \geq 0}$ , the degree distribution of the HCM with this household structure. For a vertex in the household model to have degree  $k$ , it must be in a community of size  $k$ . Furthermore, there are  $k$  of such vertices inside each community, so that

$$\hat{p}_k = \frac{kp_{k,k}}{\sum_{i=1}^{\infty} ip_{i,i}} = \frac{kCk^{-\gamma}}{\langle s \rangle} = C_2 k^{-\gamma+1}. \quad (25)$$

Thus, the degree distribution of the graph with household communities again obeys a power law but with exponent  $\tau = \gamma - 1$ , as observed in [38]. We call this phenomenon a *power-law shift*, because the edges out of a community have a smaller degree distribution than the individual edges (see Figure 3).

*b. Extremely dense communities.* In real-world networks, communities may not be complete, but a power-law shift also occurs in networks with an incomplete but extremely dense community structure. In an extremely dense community, many edges are contained in communities. Let  $e_{\text{in}}$  denote the number of edges inside a community. We assume that there exists  $\varepsilon > 0$  independent of the number of communities  $n$  such that for each community  $H$ ,

$$e_{\text{in}} \geq \varepsilon s(s-1). \quad (26)$$

In this case, every community of size  $s$  contains a positive fraction of the edges that are present in a complete graph

of the same size. Note that the household model gives  $\varepsilon = \frac{1}{2}$ .

Since the power-law shift states that the outside degrees of the communities are ‘small’, we need the outside degree of individual vertices to be small as well. Thus, we assume that there exists a  $K < \infty$  such that for all vertices

$$d_v^{(b)} \leq K s_i. \quad (27)$$

Note that this implies that  $k \leq K s^2$  for every community. Using assumptions (26) and (27) we show that a power-law shift occurs.

Suppose that the community size distribution follows a power law with exponent  $\gamma$ . Denote the cumulative degree distribution by  $P_i = \sum_{j \leq i} \hat{p}_j$ . Since the maximal inside degree of a vertex is  $s - 1$ , and by (26) the average inside degree of a vertex is greater than or equal to  $\varepsilon(s - 1)$ , at least a fraction of  $\varepsilon$  vertices in any community have inside degree at least  $\varepsilon(s - 1)$ . Thus, a vertex inside a community of size  $i/\varepsilon + 1$  has probability of at least  $\varepsilon$  to have inside degree at least  $\varepsilon(i/\varepsilon + 1 - 1) = i$ . Hence,  $1 - P_i$  is bounded from below by  $\varepsilon$  times the probability of choosing a vertex in a community of size at least  $i/\varepsilon + 1$ . The probability that a randomly chosen vertex is in a community of size  $j$  is given by  $\sum_k r_{k,j}$ . This yields

$$\begin{aligned} 1 - P_j &\geq \sum_{i \geq j} \sum_k r_{k,i/\varepsilon+1} \varepsilon \\ &= \sum_{i \geq j} \left( \frac{i}{\varepsilon} + 1 \right) \sum_k p_{k,i/\varepsilon+1} \frac{1}{\langle s \rangle} \varepsilon \\ &\approx C j^{-\gamma+2}. \end{aligned} \quad (28)$$

Furthermore, given the distribution of the community sizes,  $1 - P_j$  is maximal when all communities are complete graphs, and every vertex has  $Ks$  half-edges attached to it. Then each vertex in a community of size  $s$  has degree  $s - 1 + Ks$ . Hence, to choose a vertex with degree at least  $j$ , we have to choose a vertex inside a community of size at least  $(j + 1)/(K + 1)$ . Then

$$\begin{aligned} 1 - P_j &\leq \sum_{i \geq \frac{j+1}{K+1}} \sum_k r_{k,i} \\ &= \sum_{i \geq \frac{j+1}{K+1}} i \sum_k p_{k,i} \frac{1}{\langle s \rangle} = \frac{c}{\langle s \rangle} \left( \frac{j+1}{K+1} \right)^{-\gamma+2}. \end{aligned} \quad (29)$$

Combining (28) and (29) shows that the degree distribution follows a power law with exponent  $\tau = \gamma - 1$ . In other words, when the community size distribution of a network with extremely dense communities follows a power law with exponent  $\gamma$ , the power law of the degrees has exponent  $\tau = \gamma - 1$ .

Under a more strict assumption on the inter-community degrees

$$d_v^{(b)} \leq K, \quad (30)$$

we can also relate the power-law exponent of the intra-community degrees to the exponent of the degree distribution. Assumption (30) implies  $k \leq sK$ , and therefore if the community size distribution follows a power law with exponent  $\gamma = \tau + 1$ , then the distribution of the community outside degrees cannot have a power-law distribution with exponent smaller than  $\gamma$ . Suppose we want to construct a graph where the degree distribution follows a power law with exponent  $\tau \in (2, 3)$ . One possibility to construct such a graph is to use the CM. However, the CM with  $\tau \in (2, 3)$  has probability zero to create a simple graph. Another way to construct a graph with this degree distribution is to use the HCM with extremely dense communities of power-law size with exponent  $\tau + 1$ . The outside degrees of the communities then follow a power law with exponent at least  $\tau + 1 \geq 3$ . Since the outside degrees are paired according to the CM, the probability that the resulting graph is simple, will be larger than zero in the limit of infinite graph size. Thus, the HCM is able to construct a simple graph with exponent  $\tau \in (2, 3)$ .

Another interesting application of this power-law shift is in the critical percolation value. It is well known that the critical percolation value  $\phi_c = 0$  for a CM with  $\tau \in (2, 3)$  [28]. Section IIIB showed that for highly connected communities, the critical percolation value is entirely defined by the inter-community degrees. Since the inter-community degrees have exponent larger than 3, the HCM is able to construct random graphs with  $\tau \in (2, 3)$  and  $\phi_c > 0$ . This shows that the HCM with extremely dense communities is in another universality class than the CM.

*c. The role of hubs.* A power-law degree distribution implies the existence of hubs: nodes with a very high degree. We now show that this can conflict with assumptions (26) and (27). Since every vertex is inside a community in the HCM, the hubs also need to be assigned to some community. In these communities, hubs can have several roles, as observed in [17]. There are two possibilities, as shown in Figure 4. When most neighbors of the hub are also inside the community as in Figure 4a, then the hub is in a very large community. Assumption (26) states that most neighbors of the hub should also be connected to one another, and thus also have a high degree. However, in real-world networks this might not be realistic. For example, when one person in a social network has many friends, this does not mean that most of these friends are friends with one another. Hence, putting most neighbors of a hub inside the same community can create communities that are not dense. The other possibility (see Figure 4a) is to have only a small fraction of the neighbors inside a community. However, now the outside degree of the hub is large, which may contradict assumption (27) when the hub is in a small community. Therefore, the existence of hubs conflicts with the assumption of extremely dense communities.

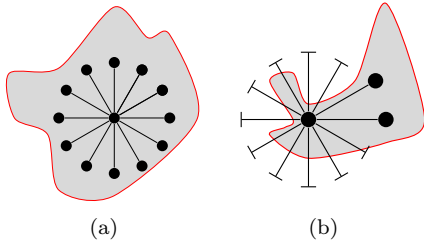


Figure 4. (a): a hub with all its neighbors completely inside a community (shaded area). (b): A hub with only a few neighbors inside the same community

Table I. Several characteristics of four different data sets

	Amazon	Gowalla	WordNet	Google
$S$ (data)	1,000	1,000	0,994	0,977
$S$ (HCM)	1,000	1,000	0,994	0,978
$S$ (CM)	0,999	0,993	0,999	0,997
$\gamma$	3,84	2,44	3,23	2,58
$\tau$	3,59	2,48	2,82	2,73
$\alpha$	0,15	0,31	0,21	0,21
$\beta$	1,14	1,18	1,28	1,24
$\gamma/\alpha - 1$	25,04	6,85	14,17	11,20

## V. REAL-WORLD NETWORKS

In this section, we apply the HCM to four different data sets: an AMAZON co-purchasing network [40], the GOWALLA social network [10], a network of relations between English words [27] and a GOOGLE web graph [25]. To extract the community structure of the networks, we use the Infomap community detection method [36], a community detection method that performs well on several benchmarks [24]. Table I shows that equation (7) identifies the size of the giant component almost perfectly, in contrast to the value calculated by the CM.

Figure 5 illustrates the power laws for these data sets. Table I presents values of the power-law exponents  $\tau$  and  $\gamma$ , estimated by the method of Clauset et al. [12]. We see that a power-law shift is less pronounced than in the stylized household model, if existing at all. This indicates that the communities in the data sets do not have the intuitive dense structure. Thus, we test assumption (26). The maximum number of edges inside a community is obtained if the community is a complete graph, in which case  $e_{\text{in}} = \frac{1}{2}s(s-1)$ . Dividing (26) by  $s(s-1)/2$  gives  $\frac{e_{\text{in}}}{s(s-1)/2} \geq 2\varepsilon$ . This fraction measures how dense a community is. Figure 6 plots  $s$  against the average value of  $2e_{\text{in}}/(s^2 - s)$ . For all networks, this fraction is not independent of  $s$ . Larger communities are less dense than smaller communities. Therefore, the large communities do not satisfy the intuitive picture of an extremely densely connected subset, even though the density within communities is much higher than that in the entire network. This is a similar observation as in [25], where the

authors discover that most real-world networks have a strongly connected core, which consists of several interconnected communities that are hard to distinguish. The core is connected to the periphery, some isolated, densely connected small communities. This structure could explain the dependence of the density of the communities on  $s$ . The large communities that are not very dense, are part of the core, whereas the small communities are the more isolated parts of the network. Another interesting property of the community structures in Figure 6 is the power-law relation between the community sizes and their densities,  $e_{\text{in}} \approx cs^{\alpha+1}$ . In assumption (26), we assume that  $\alpha = 1$ . However, Table I shows that the example data sets have  $\alpha < 1$ . For this reason, we replace (26) by

$$e_{\text{in}} \geq \varepsilon s(s-1)^\alpha. \quad (31)$$

Now (29) still holds, but (28) needs to be modified. The average inside degree of a vertex now is  $\varepsilon(s-1)^\alpha$ . Since the maximum inside degree is  $s-1$ , there are at least  $\varepsilon(s-1)^{\alpha-1}$  vertices of degree at least  $\varepsilon(s-1)^\alpha$ . A similar analysis as (28) yields

$$\begin{aligned} 1 - P_j &\geq \sum_{i \geq j} \sum_k r_{k, (i/\varepsilon)^{(1/\alpha)+1}} \varepsilon (i/\varepsilon)^{(\alpha-1)/\alpha} \\ &= \frac{\varepsilon}{\langle s \rangle} \sum_{i \geq j} \left( \left( \frac{i}{\varepsilon} \right)^{(1/\alpha)} + 1 \right)^{-\gamma+1} (i/\varepsilon)^{(\alpha-1)/\alpha} \\ &\approx C j^{-\gamma/\alpha+2}. \end{aligned} \quad (32)$$

Together with (29), this shows that the exponent  $\tau$  of the degree distribution satisfies  $\tau \in [\gamma - 1, \frac{\gamma}{\alpha} - 1]$ . Table I shows several values of  $\tau$ ,  $\gamma$  and  $\gamma/\alpha - 1$ . We see that indeed  $\tau \in [\gamma - 1, \frac{\gamma}{\alpha} - 1]$  in the example data sets. However, the interval may be quite wide.

We next test assumption (27). Interestingly, Figure 7 shows a power-law relationship between  $k$  and  $s$ , of the form  $k \approx s^\beta$ . If  $k \leq Ks^2$  would hold, then  $\beta \leq 2$ , whereas the more strict assumption (30) would imply  $\beta < 1$ . Table I shows that the example data sets all have  $1 < \beta < 2$ . Therefore, the more strict assumption (30) does not hold, but (27) does hold. Thus, large communities have very large outside degrees.

## VI. CONCLUSIONS AND OUTLOOK

We have introduced the Hierarchical Configuration Model (HCM) as a random graph model that can describe both realistic degree distributions and an arbitrary community structure, while remaining analytically tractable in the large-graph limit. Our analysis of the HCM has revealed several properties. The condition for a giant component to emerge in the HCM is completely determined by properties of the macroscopic configuration model at the level of communities and therefore not affected by the precise structure or size of communities.



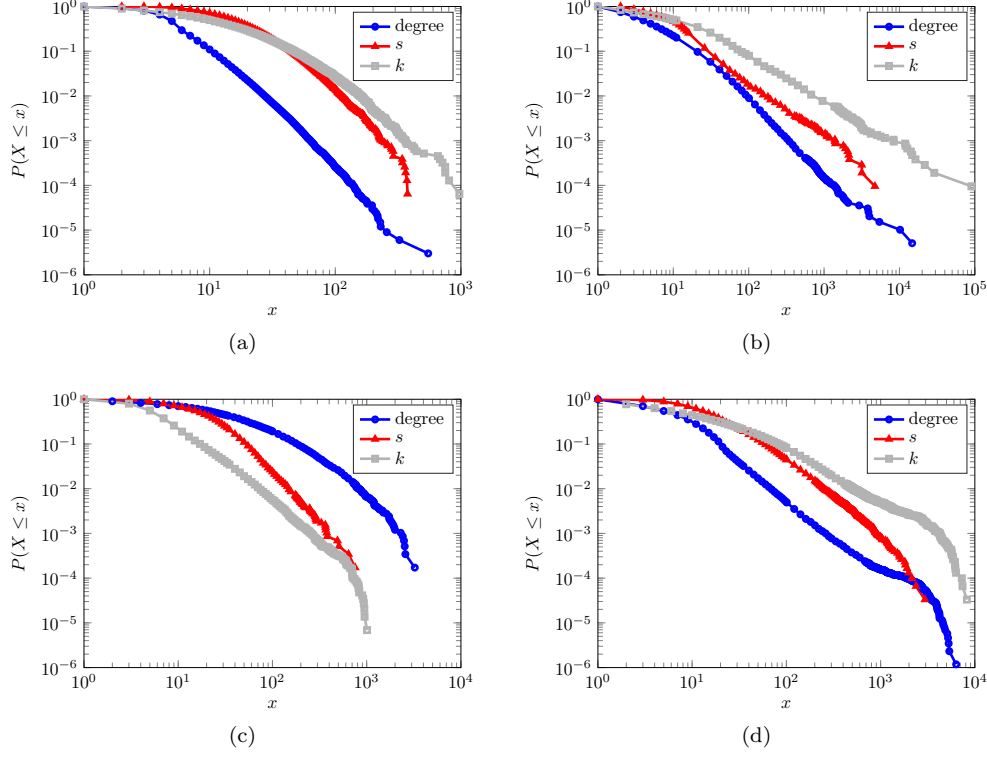


Figure 5. Power-law relations in real-world networks. a) AMAZON co-purchasing network, b) GOWALLA social network, c) English word relations, d) GOOGLE web graph. Both the degrees and the community sizes  $s$  follow a power law, as well as the inter-community degrees  $k$ .

The size of the giant component, however, strongly depends on the joint probability distribution describing the size of a community and its outside degree. Under bond percolation, communities may either increase or decrease the critical percolation value compared to a configuration model with the same degree sequence.

For the prototypical case of extremely dense communities, we show that a power-law degree distribution with exponent  $\tau$  implies a power-law distribution for the community sizes with exponent  $\gamma = \tau + 1$ . Real-world networks, however, rarely possess an extremely dense community structure [25].

Studying the HCM allows us to observe two previously unobserved power-law relations in several real-world networks. The relation between the number of edges inside a community  $e_{in}$  and the community sizes  $s$  follows a power law of the form  $e_{in} \propto s^{1+\alpha}$ . The second power-law relation is between the number of edges going out of a community  $k$  and the community sizes:  $k \propto s^\beta$ . The data sets that were studied in this paper had  $1 < \beta < 2$  and  $\alpha < 1$ . Combined, the two power-law relations improve our understanding of the community structure in the data sets. Large communities are not extremely densely connected, and have a large number of edges going out of the community per vertex. Smaller communities are dense, and vertices in the community have only a few edges going out of the community. Our intuitive picture of extremely

densely connected communities thus only holds for the small communities in a network, the larger communities do not fit into this picture. The observation that large communities are not extremely dense may be a consequence of not allowing for overlapping communities. In case of several overlapping communities, community detection algorithms may merge these communities into one large community. As a consequence, this large community will be far from extremely dense. In the case of overlapping communities, many networks still display a power-law community size distribution [34]. It would be interesting to investigate the relation between the exponent of the degree distribution and the community size distribution when communities are allowed to overlap. Further research could also study how the denseness of the communities and the number of edges out of the communities are related to the community sizes in the case of overlapping communities.

Both power-law relations are observational, and therefore depend on the Infomap community detection algorithm. It is also possible to use other community detection algorithms to investigate these power-law relations. We found that when using the Louvain community detection algorithm [6], the power-law relations still hold. The estimates for the exponents  $\alpha$  and  $\beta$  however did change. This can be explained by the fact that the Louvain method finds larger communities in general, which

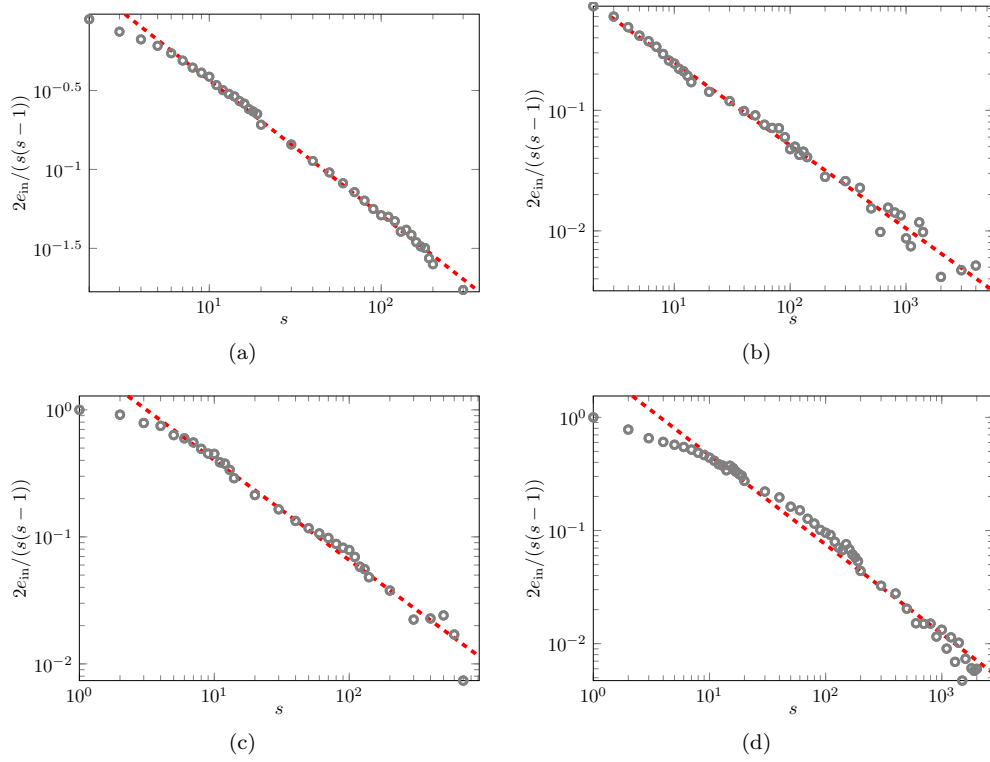


Figure 6. The denseness of a community  $2e_{\text{in}}/(s^2 - s)$  has a power-law relation with the community size  $s$ . a) AMAZON co-purchasing network, b) GOWALLA social network, c) English word relations, d) GOOGLE web graph.

are therefore less dense.

The power-law exponent of the degree distribution  $\tau$  is known to influence the behavior of various processes on random graphs, for example percolation or epidemic models. Furthermore, mean distances in random graphs are different for  $\tau \in (2, 3)$ , or  $\tau > 3$  [19, 20]. Our results suggest that for networks with a community structure it is not clear whether the behavior of these processes can be explained by  $\tau$ , or the exponent of the community degrees, this remains open for further research. The results on the power-law shift suggest that this may depend on the density of the communities, which is characterized by the exponent  $\alpha$ .

The HCM keeps all edges inside the communities, while rewiring the inter-community edges. Instead of fixing the precise internal community structure, one could also randomize the edges inside communities as in a CM. This model was introduced as the modular random graph [37], a random graph with a given degree distribution and modularity. The focus in [37] is on the algorithmic construction of the modular random graph, not on the analytical properties of this model. The analytic study of the modular random graph is worthwhile to pursue. From the present work, it is clear that the analysis of the giant component remains the same as for the HCM, at least when the communities are likely to be connected, so the precise details of the internal community structure can be safely ignored. However, we have also seen that

the internal community structure does become important when considering the critical percolation threshold, and in this case the analysis of the HCM does not carry over to the modular random graph.

### Acknowledgement

This work is supported by NWO TOP grant 613.001.451 and by the NWO Gravitation Networks grant 024.002.003. The work of RvdH is further supported by the NWO VICI grant 639.033.806. The work of JvL is further supported by an NWO TOP-GO grant and by an ERC Starting Grant.

### Appendix A: Influence of communities on percolation

In this section, we study two examples of community structures. The first example decreases the critical percolation value when comparing the HCM with the CM with the same degree distribution. The second example increases the critical percolation value when comparing the HCM with the CM.

As an example of a graph that decreases  $\phi_c$ , consider a network where with probability  $\zeta$  a community is given by  $H_1$ : a path of  $l$  vertices, with an outgoing half-edge

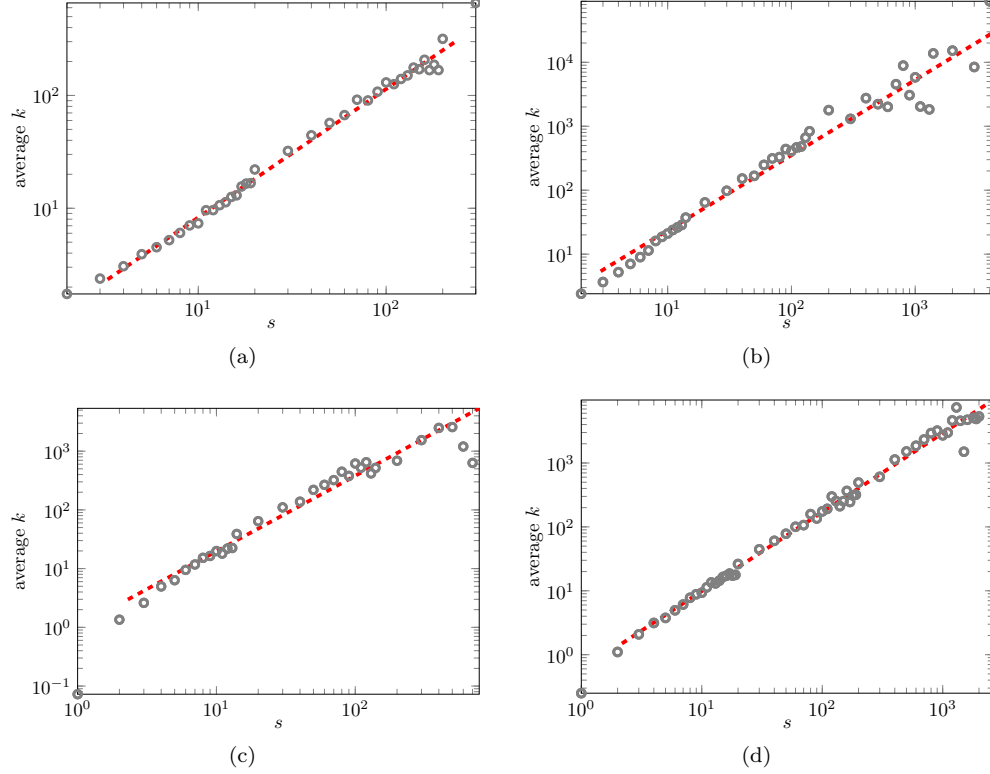


Figure 7. The outside degree of a community  $k$  follows a power-law relation with the community size  $s$ . a) AMAZON co-purchasing network, b) GOWALLA social network, c) English word relations, d) GOOGLE web graph.



Figure 8. A line community with  $l = 5$

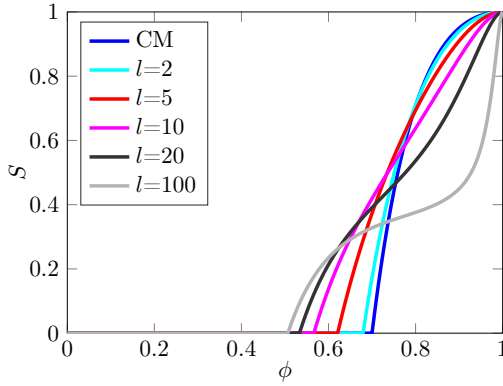


Figure 9. Size of largest percolating cluster  $S$  for HCM with line communities of length  $l$ , while the degree distribution remains the same. The line communities decrease the critical percolation value.

at each end of the path as illustrated in Figure 8. With probability  $1-\zeta$  the community is  $H_2$ : a vertex with three outgoing half-edges. Then  $\langle k \rangle = 2\zeta + 3(1-\zeta) = 3-\zeta$  and  $f(H_1, v, 2, \phi) = \phi^{l-1}$  for all  $v \in H_1$ . In  $H_2$  there is no

percolation inside the community, hence  $f(H_2, v, 3, \phi) = 1$ . Using (24) yields

$$\phi_c = \frac{3-\zeta}{2\zeta\phi_c^{l-1} + 6(1-\zeta)}, \quad (\text{A1})$$

hence  $2\zeta\phi_c^l + 6(1-\zeta)\phi_c - 3 + \zeta = 0$ .

Consider this version of the HCM with degree distribution  $p_2 = 1 - p_3 = \frac{2}{3}$ . We keep the degree distribution the same, while the length  $l$  of the line communities  $H_1$  changes. Then if  $l$  increases,  $\zeta$  decreases to keep the degree distribution the same. Using (22) and (23), we find the size of the giant component, as depicted in Figure 9. Note that  $\phi_c$  decreases with  $l$ , this community structure ‘helps’ the diffusion process. This can be explained by the fact that there will be fewer line communities if  $l$  increases. Then most vertex communities are connected to one another, which decreases the value of  $\phi_c$ . Interestingly, the size of the giant component in the HCM is non-convex in  $\phi$ . This is different than in the CM, where the percolation curves are typically convex. The non-convex shapes can be explained intuitively. As the lines get longer, there are fewer and fewer of them, since the degree distribution remains the same. Hence, if  $l$  is large, there are only a few very long lines. These lines have  $\phi_c = 1$ . The other vertices are of degree 3, connected as the CM. Since there are only a few lines, most vertices of degree 3 will be paired to one another. The

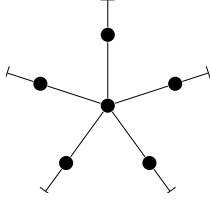


Figure 10. A star-shaped community with  $l = 5$

critical value for percolation on a CM with only vertices of degree 3 is  $\frac{1}{2}$ . Therefore, for large  $l$  we see the vertices of degree 3 appearing in the giant component as  $\phi = 0.5$ , and the vertices in the line communities as  $\phi = 1$ .

An example of a network that inhibits the diffusion process is a CM with intermediate vertices [26], where every edge is replaced by two edges and a vertex in between them. This is the same as the HCM with star-shaped communities: one vertex that is connected to  $l$  other vertices. Each of the  $l$  other vertices has outside degree one (Figure 10). In this example, we consider a HCM where all communities have the same star-size  $l$ , so that  $\langle k \rangle = l$ . After percolation, the connected component of an end point of the star can link to other outgoing edges only if the edge to the middle vertex is present. If this edge is present, the number of half-edges to which

the vertex is connected is binomially distributed:

$$f(H, v, k, \phi) = \phi \binom{l-1}{k-1} \phi^{k-1} (1-\phi)^{l-k} \quad k \geq 2. \quad (\text{A2})$$

Hence by (24),

$$\begin{aligned} \phi_c &= \frac{l}{l\phi_c \sum_{k \geq 1} k \phi_c^k (1-\phi_c)^{l-k-1} \binom{l-1}{k}} \\ &= \frac{1}{(l-1)\phi_c^2}, \end{aligned} \quad (\text{A3})$$

so that  $\phi_c = (l-1)^{-1/3}$ .

We next consider a CM with the same degree distribution. Figure 11 shows the size of the giant component for different values of  $l$  for both the HCM and the CM. The HCM with star communities has a higher critical percolation value than the corresponding CM. Intuitively, this can be explained by the fact that all vertices with a high degree are ‘hidden’ behind vertices of degree 2, whereas in the CM, vertices with degree  $l$  may be connected to one another. However, as  $\phi$  increases, at some point, the star communities make the giant component larger.

Combined with the previous example, we see that adding communities may lead to a higher critical percolation value or a lower one. Furthermore, the size of the giant component may be smaller or larger after adding communities.

- 
- [1] R. Albert, H. Jeong, and A.-L. Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.
  - [2] F. Ball, D. Sirl, and P. Trapman. Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. in Appl. Probab.*, 41(3):765–796, 09 2009.
  - [3] F. Ball, D. Sirl, and P. Trapman. Analysis of a stochastic SIR epidemic on a random network incorporating household structure. *Mathematical Biosciences*, 224(2):53–73, 2010.
  - [4] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(14):69–77, 2000.
  - [5] S. Bhamidi, R. v. d. van der Hofstad, and G. Hooghiemstra. First passage percolation on random graphs with finite mean degrees. *The Annals of Applied Probability*, 20(5):pp. 1907–1965, 2010.
  - [6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
  - [7] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Phys. Rev. E*, 70:056122, Nov 2004.
  - [8] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European*

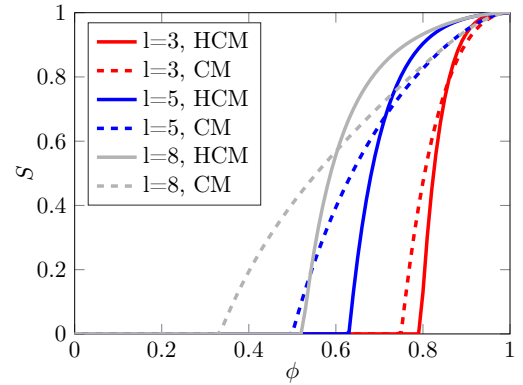


Figure 11. Size of largest percolating cluster  $S$  for the HCM with star communities of size  $l$ , compared to the CM. The star communities increase the critical percolation value.

- Journal of Combinatorics*, 1(4):311–316, 1980.
- [9] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85(25):5468, 2000.
- [10] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD in-*

- ternational conference on Knowledge discovery and data mining, pages 1082–1090. ACM, 2011.
- [11] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004.
  - [12] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
  - [13] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685, Apr 2001.
  - [14] E. Coupechoux and M. Lelarge. How clustering affects epidemics in random networks. *Adv. in Appl. Probab.*, 46(4):985–1008, 12 2014.
  - [15] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.
  - [16] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
  - [17] R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.
  - [18] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68:065103, Dec 2003.
  - [19] R. van der Hofstad, G. Hooghiemstra, and P. Van Mieghem. Distances in random graphs with finite variance degrees. *Random Structures & Algorithms*, 27(1):76–123, 2005.
  - [20] R. van der Hofstad, G. Hooghiemstra, and D. Znamenski. Distances in random graphs with finite mean and infinite variance degrees. *Electron. J. Probab.*, 12:no. 25, 703–766, 2007.
  - [21] R. van der Hofstad, J. S. H. van Leeuwen, and C. Stegehuis. Hierarchical configuration model. Preprint 2015.
  - [22] S. Janson. On percolation in random graphs with given vertex degrees. *Electron. Journal of Probability*, 14:86–118, 2009.
  - [23] B. Karrer and M. E. J. Newman. Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E*, 82:066118, Dec 2010.
  - [24] A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.
  - [25] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.
  - [26] N. Litvak and R. van der Hofstad. Uncovering disassortativity in large scale-free networks. *Phys. Rev. E*, 87:022801, Feb 2013.
  - [27] G. Miller and C. Fellbaum. Wordnet: An electronic lexical database, 1998.
  - [28] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.
  - [29] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
  - [30] M. E. J. Newman. Random graphs with clustering. *Phys. Rev. Lett.*, 103(5):058701, July 2009.
  - [31] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, 2010.
  - [32] M. E. J. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66:035101, Sep 2002.
  - [33] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, 2001.
  - [34] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
  - [35] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.
  - [36] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
  - [37] P. Sah, L. O. Singh, A. Clauset, and S. Bansal. Exploring community structure in biological networks with random graphs. *BMC Bioinformatics*, 15(1):220, 2014.
  - [38] P. Trapman. On analytical approaches to epidemics on networks. *Theoretical Population Biology*, 71(2):160 – 173, 2007.
  - [39] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *Phys. Rev. E*, 65(6):066130, 2002.
  - [40] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.